4M1-2

工学システムの間欠的センサデータのための動的混合モデル

Dynamic Mixture Models for Intermittent Sensor Data of Engineering Systems

武石 直也[*1]
Naoya Takeishi

矢入 健久[*1]
Takehisa Yairi

[*1] 東京大学大学院 工学系研究科
School of Engineering, the University of Tokyo

Engineering systems such as industrial plants, cars, and artificial satellites, are equipped with many sensors, which generate a huge amount of time-series data. In practice, analyzing and modeling those time-series data are often challenging when they are intermittent, that is, unevenly spaced and temporally sparse. This work proposes a generative model for the intermittent multivariate time-series, and it is a hierarchical Bayesian mixture model in which priors of mixture assignment evolve over time, with whole data partitioned into several groups by temporal proximity.

1. 背景

工業プラント、自動車、人工衛星などの工学システムから得られるセンサデータは、故障検出や運用最適化などのための情報として重要である。そのようなセンサデータは、多くの場合に時系列データとして得られる。したがって、モデル化のアプローチとしては、隠れマルコフモデルや状態空間モデルなどの時系列モデルを用いることが一般的である。しかし、一般的な時系列モデルがただちに適切に用いられる「きれいな」センサデータが得られることはまれであろう。そのため、各種の前処理が必要となるが、単純な前処理では十分に対処できなかったり、モデルの推定量に大きなバイアスが生じたりすることがある。

本稿は、工学システムから得られるセンサデータからモデルを推定して利用する際の、実際的な困難のひとつに対処する生成モデル的アプローチを提案する。本稿で対象とするのは、時系列データとしての「間欠性」である（図 1）。定性的な定義にとどまるが、時系列データとしての間欠性として次の性質を想定する。

① サンプリング周期が一定でない (unevenly spaced): 観測が得られる周期が途中で変化する性質
② 時間的に疎 (temporally sparse): 長期間の欠測が不規則に存在する性質

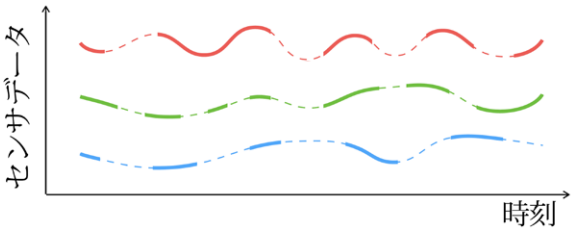

図 1 間欠的時系列データのイメージ。実線が観測部分を、破線が欠測部分を示す。

センサデータの間欠性は、データ解析・利用の現場においてしばしば課題となる性質であろう。直感的な解決策のひとつは、時系列データの補間（内挿）である。しかし、補間によってモデルの推定量にバイアスが生じてしまうし、ノイズの大きいデータではよい補間は難しい。また、データの時間的疎性が限定的な場合には、フィルタリングやスムージングの際に欠測時刻をスキップしたり、連続時間モデルを用いたりすることが可能である。しかし、そのようなモデルは高周波成分に焦点をおいており、長期間の欠測がしばしば入るようなデータに対応できない。

サンプリング周期が一定でない時系列データに関しては、基本的なオペレータ、モデルや周期性などについて研究されてきた（例えば、[Erdogan 05, Zumbach 01, Jones 85, Foster 96]など）。ところが、多くは単変数で定常な時系列に関するモデルや手法であって、多変数で非定常性の強い工学システムのセンサデータに対して適用しにくい。

間欠的なセンサデータを扱うにあたり、実用面で強力な方法は、時系列性を無視してしまうことであろう。各観測が独立に同分布から生成されたと見なすことによって、適用できるモデルの幅が広がるうえ、モデル推定も簡単になる。特に、興味のある情報が時間的相関ではなく空間的相関（センサ間の相関）にあるような場合には、この方法は特に有効である。ところが当然ながら、この方法では時間的な情報を失ってしまう。

さて本稿では、時間的間欠性を念頭においた生成モデルを用いて間欠的なセンサデータをモデル化し、各種タスクに利用することを考える。提案するモデルは、混合比率が時間的に変化する階層ベイズ的な混合モデルである。特に、混合比率の時間変化を、すべての観測毎ではなく一定範囲の観測毎に考えることで、時間的間欠性に対処する。以降、第 2 節に提案モデルの詳細を、第 3 節に簡単な実験の結果を示す。

2. 提案手法

提案するアプローチは任意の混合モデルに対して適用可能であるが、ここでは簡単さと有用性を鑑みて、確率的主成分分析の混合モデル[Tipping 99]を拡張することを考える。混合確率的主成分分析については、文献[Tipping 99]を参照されたい。

2.1 データのグループ分割

時間的間欠性に対処するため、時間的に近い観測ごとに、全データを複数のグループに分割する。すなわち、観測$\mathbf{y}_m$が行われた時刻を$h(m)$として、$s_t \leq h(m) < s_{t+1}$のとき、観測$\mathbf{y}_m$はt番目のグループに属するとする。また、今回は簡単のため、各グループの時間幅$\Delta s_t \equiv s_{t+1} - s_t$はすべての$t$について一定とする。ただし、各グループに属する観測の数は一定ではな

連絡先：takeishi@ailab.t.u-tokyo.ac.jp

いことに注意されたい。また、このΔs_tはハイパーパラメタとして扱う。Δs_tを小さくとれば高周波成分に、Δs_tを大きくとれば低周波成分に着目することになるため、タスクに応じて設定する必要がある。

2.2 生成モデル

提案する生成モデルは、次のようなものである。

t番目のグループについて、

1. 混合比率の事前分布: $\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1} \sim \mathrm{Normal}(\boldsymbol{\eta}_{t-1}, \Lambda)$
2. グループ中の全観測 $n = 1, \dots, N_t$ について、
   A) 混合割当: $\mathbf{z}_{t,n}|\boldsymbol{\eta}_t \sim \mathrm{Categorical}(\mathrm{softmax}(\boldsymbol{\eta}_t))$
   B) 隠れ因子: $\mathbf{x}_{t,n,k} \sim \mathrm{Normal}(\mathbf{0}, \mathrm{I})$, where $k = z_{t,n}$
   C) 観測: $\mathbf{y}_{t,n}|\mathbf{z}_{t,n}, \mathbf{x}_{t,n,k} \sim \mathrm{Normal}(\mathrm{L}_k\mathbf{x}_{t,n,k} + \mathbf{b}_k, \Psi_k)$

ただし、N_tはt番目のグループに属する観測の数である。また、Λは混合比率の事前分布の時間発展分散であり、$\mathrm{L}, \mathbf{b}, \Psi$は主成分分析のパラメタである。さらに、$k = 1, \dots, K$であり、$K$は混合モデルの混合要素数を表すハイパーパラメタである。

2.3 推論と学習

隠れ変数の推論やモデルパラメタの学習のために、今回は変分ベイズ推論を用いた EM アルゴリズムを採用した。必要な計算の詳細は文献[Takeishi 16]を参照のこと。

3. 実験

提案モデルの有用性を示すための実験のひとつとして、時系列データのノイズ除去(デノイジング)を行った。この節ではその設定と結果を示す。

3.1 データ生成

線形状態空間モデルで人工的にデータを生成する。生成に使用したモデルは次の通りである。

$$\mathbf{y}_m = \bar{\mathbf{y}}_m + \boldsymbol{e}_m$$
$$\boldsymbol{e}_m \sim \mathrm{Normal}(\mathbf{0}, \nu^2\mathrm{I})$$
$$\mathbf{x}_m = \begin{bmatrix} -0.6 & 0.2 \\ -0.1 & 0.5 \end{bmatrix} \mathbf{x}_{m-1} + \mathbf{w}_m$$
$$\bar{\mathbf{y}}_m = \mathrm{A}\mathbf{x}_m$$

ただし、Aは8×2行列である。すなわち、最終的に生成する時系列 $\{\mathbf{y}_m\}$ $(m = 1, \dots, 5000)$ は、8 次元時系列となる。また、$\mathbf{w}_m$は、平均0、分散0.1の正規分布に従うノイズである。

時間的間欠性を模すため、上記のモデルで作成した 8 次元時系列に対してランダムに欠測を加えた。欠測の割合は、0%(欠測なし)、20%、40%、60%、80%の 5 種類を試した。

3.2 ノイズ除去

ノイズ除去のタスクは、時系列$\{\mathbf{y}_m\}$が得られたときに時系列$\{\bar{\mathbf{y}}_m\}$を推定することである。上述のように生成した時系列データのうち最初の 3000 時刻を訓練、次の 1000 時刻をバリデーション、最後の 1000 時刻をテストに用いて、ベースラインとした手法のノイズ除去性能と提案モデルによるノイズ除去性能を比較した。

表 1 に、ベースライン手法(隠れマルコフモデル=HMM、線形状態空間モデル=LDS、混合確率主成分分析=MPPCA)および提案モデル(Dynamic Grouped MPPCA, DGMPPCA)のノイズ除去性能を、推定結果と真値との RMS 誤差で示した。HMM と LDS のモデル推定時には、欠測部をスキップしてフィルタリング・スムージングを行っている。データが LDS によって生成されているため、LDS のデノイジング性能がよいと期待されるが、欠測の割合が多くなるにつれて性能が悪化した。一方、提案手法では時間的間欠性を念頭においたモデル化を行っているため、欠測の割合が多くなっても性能の悪化が見られなかった。

表 1 ノイズ除去性能(RMS error)

欠測	HMM	LDS	MPPCA	DGMPPCA
0%	2.80×10^{-1}	$\mathbf{1.29 \times 10^{-1}}$	1.77×10^{-1}	$\mathbf{1.24 \times 10^{-1}}$
20%	3.04×10^{-1}	$\mathbf{1.30 \times 10^{-1}}$	1.69×10^{-1}	$\mathbf{1.26 \times 10^{-1}}$
40%	3.15×10^{-1}	1.39×10^{-1}	1.70×10^{-1}	$\mathbf{1.25 \times 10^{-1}}$
60%	3.19×10^{-1}	1.46×10^{-1}	1.98×10^{-1}	$\mathbf{1.24 \times 10^{-1}}$
80%	3.45×10^{-1}	2.13×10^{-1}	1.92×10^{-1}	$\mathbf{1.25 \times 10^{-1}}$

4. まとめ

工学システムのセンサデータは、その時間的間欠性から一般的な時系列データの適用が難しいことがある。そこで本稿では、混合比率が時間的に変化する階層ベイズ的混合モデルを間欠的時系列データに対して適用することを提案した。また、提案モデルの有用性を確認するため、簡単なデノイジングの実験を行い、その結果を示した。

謝辞

本研究の遂行にあたり、宇宙航空開発研究機構の西村尚樹氏、中島佑太氏、高田昇氏から、データの提供や有用な議論をいただきました。

参考文献

[Erdogan 05] Erdogan, E.: Statistical models for unequally spaced time series, In *Proceedings of the 5th SIAM International Conference on Data Mining*, pp. 626‐630 (2005)

[Zumbach 01] Zumbach, G. and Müller, U.: Operators on inhomogeneous time series, *International Journal of Theoretical and Applied Finance*, vol. 4, pp. 147‐177 (2001)

[Jones 85] Jones, R.: Time series analysis with unequally spaced data, in *Handbook of Statistics*, vol. 5 (1985)

[Foster 96] Foster, G.: Wavelets for period analysis of unevenly sampled time series. *The Astronomical Journal*, vol. 112, pp. 1709‐1729 (1996)

[Tipping 99] Tipping, M. E. and Bishop, C. M.: Mixtures of probabilistic principal component analysers, *Neural Computation*, vol. 11, pp. 443‐482 (1999)

[Takeishi 16] Takeishi, N., Yairi, T., Nishimura, N., Nakajima, Y. and Takata, N.: Dynamic grouped mixture models for intermittent multivariate sensor data, in Proceedings of The 20th Pacific Asia Conference on Knowledge Discovery and Data Mining, to be published (2016)